# MINING ONLINE MUSIC LISTENING TRAJECTORIES

**Flavio Figueiredo**[1] **Bruno Ribeiro**[2] **Christos Faloutsos**[3]
**Nazareno Andrade**[4] **Jussara M. Almeida**[5]
[1] IBM Research - Brazil [2] Purdue University [3] Carnegie Mellon University
[4] Universidade Federal de Campina Grande [5] Universidade Federal de Minas Gerais

## ABSTRACT

Understanding the listening habits of users is a valuable undertaking for musicology researchers, artists, consumers and online businesses alike. With the rise of Online Music Streaming Services (OMSSs), large amounts of user behavioral data can be exploited for this task. In this paper, we present SWIFT-FLOWS, an approach that models user listening habits in regards to how user attention transitions between artists. SWIFT-FLOWS combines recent advances in trajectory mining, coupled with modulated Markov models as a means to capture both how users switch attention from one artist to another, as well as how users fixate their attention in a single artist over short or large periods of time. We employ SWIFT-FLOWS on OMSSs datasets showing that it provides: (1) semantically meaningful representation of habits; (2) accurately models the attention span of users.

## 1. INTRODUCTION

Is it possible to create expressive yet succinct representations of individuals' music listening habits? Are there common patterns on how music is listened to across different genres and different artists that have highly different popularity? For a long time such questions have attracted the attention of researchers from different fields. In the fields of psychology and musicology [10, 20, 21], researchers exploit musical preferences to study social and individual identity [20], mood regulation [23], as well as the underlying factors of preferences [21]. Computer scientists are also tackling such questions as they become central to develop music recommender systems [3, 4, 7].

With the rise of Online Music Streaming Services (OMSSs) over the last decade, large datasets of user [1] behavior can be used to shed light on questions like the ones above. More specifically, digital traces of the listening habits of individuals are readily available to researchers.

In this paper, we focus on the online listening habits of users as trajectories [7] (or trails [24]). Given that a user, $u$, listens to music by switching attention between different artists, a trajectory captures the sequence of artists or songs visited by a user when listening to music. The main contribution of this paper is to present the SWIFT-FLOWS [2] model, a general technique designed to study user trajectories in OMSSs. We tackle several challenges that stem from the complexity of user behavior, such as:

(a) *Asynchronous users with mixed but similar behavior*: Users that consume music from a set of artists will not start their playlists at the same time or listen to songs in the same order.
(b) *Repeated consumption*: Users tend to listen to artists in bursts, more than what one would expect at random in a shuffled playlist.
(c) *Biased Observations & Small Subpopulations*: User behavior datasets are naturally sparse and biased towards more popular artists. Nevertheless, we still want to be able to analyze underrepresented subpopulations of users and artists.

SWIFT-FLOWS effectiveness is evaluated in large datasets, with results showing that SWIFT-FLOWS: (1) captures semantically meaningful representation of artist transitions; (2) accurately models the attention span of users.

## 2. RELATED WORK

Understanding the listening habits of individuals has attracted interest from different research fields. Among other problems, musicologists and social psychologists have looked into the latent factors that explain musical preferences [20, 21], factors that affect listener experience (e.g., Music itself, Situational Factors and the Listener him/herself) [10], as well as the relationships between musical imagination and human creativity [10].

Regarding the material methods listeners exploit to listen to music, Nowak [16] discussed the social-material relations of music consumption. The authors conclude that even the same user still relies on multiple forms of listening to music (e.g., legal and illegal downloading, streaming services, CDs, etc). These various forms of consumption were also discussed by Bellogin *et al.* [1]. Here, the au-

[1] Since our case study is on Online Music Streaming Services (OMSSs), we use the terms users and listeners interchangeably.

[2] Switch and Fixation Trajectory Flows

Information Sciences 349–350 (2016) 172–187

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# TrendLearner: Early prediction of popularity trends of user generated content

Flavio Figueiredo*, Jussara M. Almeida, Marcos A. Gonçalves, Fabricio Benevenuto

*Department of Computer Science, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, CEP 31270-010, Belo Horizonte - MG, Brazil*

ARTICLE INFO

*Article history:*
Received 7 June 2015
Revised 6 December 2015
Accepted 12 February 2016
Available online 26 February 2016

*Keywords:*
Popularity
Trends
Classification
Social media
UGC
Prediction

ABSTRACT

Predicting the popularity of user generated content (UGC) is a valuable task to content providers, advertisers, as well as social media researchers. However, it is also a challenging task due to the plethora of factors that affect content popularity in social systems. Here, we focus on the problem of predicting the popularity *trend* of a piece of UGC (object) *as early as possible*. Unlike previous work, we explicitly address the inherent tradeoff between prediction accuracy and remaining interest in the object after prediction, since, to be useful, accurate predictions should be made *before* interest has exhausted. Given the heterogeneity in popularity dynamics across objects, this tradeoff has to be solved on a per-object basis, making the prediction task harder. We tackle this problem with a novel two-step learning approach in which we: (1) extract popularity trends from previously uploaded objects, and then (2) predict trends for newly uploaded content. Our results for YouTube datasets show that our classification effectiveness, captured by F1 scores, is 38% better than the baseline approaches. Moreover, we achieve these results with up to 68% of the views still remaining for 50% or 21% of the videos, depending on the dataset.



## 1. Introduction

The success of Internet applications based on user generated content (UGC)[1] has motivated questions such as: How does content popularity evolve over time? What is the potential popularity a piece of content will achieve after a given time period? How can we predict popularity evolution of a particular piece of UGC? For example, from a system perspective, accurate popularity predictions can be exploited to build more cost-effective content organization and delivery platforms (e.g., caching systems, CDNs). They can also drive the design of better analytic tools, a major segment nowadays [20,34], while online advertisers may benefit from them to more effectively place contextual advertisements. From a social perspective, understanding issues related to popularity prediction can be used to better understand the human dynamics of consumption. Moreover, being able to predict popularity on an automated way is crucial for marketing campaigns (e.g. created by activists or politicians), which increasingly often use the Web to influence public opinion.

* Corresponding author. Tel.: +55 31 3409 5860.
*E-mail addresses:* flaviov@dcc.ufmg.br (F. Figueiredo), jussara@dcc.ufmg.br (J.M. Almeida), mgoncalv@dcc.ufmg.br (M.A. Gonçalves), fabricio@dcc.ufmg.br (F. Benevenuto).

[1] YouTube, Flickr, Twitter, and so forth.

http://dx.doi.org/10.1016/j.ins.2016.02.025
0020-0255/

# TribeFlow: Mining & Predicting User Trajectories

Flavio Figueiredo[1,2], Bruno Ribeiro[4,5], Jussara Almeida[3], Christos Faloutsos[5]
[1]UFCG - Brazil, [2]IBM Research - Brazil, [3]UFMG - Brazil, [4]Purdue University, [5]Carnegie Mellon University
{flaviov,jussara}@dcc.ufmg.br, ribeiro@cs.purdue.edu, christos@cs.cmu.edu

## ABSTRACT

Which song will Smith listen to next? Which restaurant will Alice go to tomorrow? Which product will John click next? These applications have in common the prediction of user trajectories that are in a constant state of flux over a hidden network (e.g. website links, geographic location). Moreover, what users are doing now may be unrelated to what they will be doing in an hour from now. Mindful of these challenges we propose TribeFlow, a method designed to cope with the complex challenges of learning personalized predictive models of non-stationary, transient, and time-heterogeneous user trajectories. TribeFlow is a general method that can perform next product recommendation, next song recommendation, next location prediction, and general arbitrary-length user trajectory prediction without domain-specific knowledge. TribeFlow is more accurate and up to $413\times$ faster than top competitors.

## Keywords

User Trajectory Recommendation; Latent Environments;

## 1. INTRODUCTION

Web users are in a constant state of flux in their interactions with products, places, and services. User preferences and the environment that they navigate determine the sequence of items that users visit (links they click, songs they listen, businesses they visit). In this work we refer to the sequence of items visited by a user as the user's trajectory. Both the environment and user preferences affect such trajectories. The underlying navigation environment may change or vary over time: a website updates its design, a suburban user spends a weekend in the city. Similarly, user preferences may also vary or change over time: a user has different music preferences at work and at home, a user prefers ethnic food on weekdays but will hit all pizza places while in Chicago for the weekend.

The above facts result in user trajectories that over multiple time scales can be non-stationary (depend on wall clock times), transient (some visits are never repeated), and time-heterogeneous (user behavior changes over time); please refer to Section 5 for examples. Unfortunately, mining non-stationary, transient, and time-heterogeneous stochastic processes is a challenging task. It would be easier if trajectories were stationary (behavior is independent of wall clock times), ergodic (visits are infinitely repeated), and time-homogeneous (behavior does not change over time).

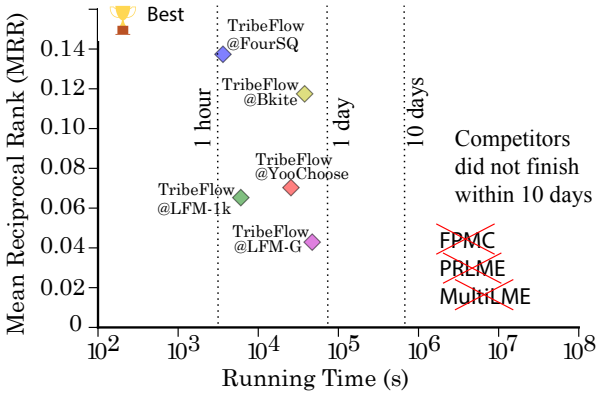



**Figure 1: TribeFlow is at least an order of magnitude faster than state-of-the-art methods for next-item predictions.**

In this work we propose TribeFlow to tackle the problem of mining and predicting user trajectories. TribeFlow takes as input a set of users and a sequence items they visit (user trajectories), including the timestamps of these visits if available, and outputs a model for personalized next-item prediction (or next $n > 1$ items). TribeFlow can be readily applied to personalized trajectories from next check-in recommendations, to next song recommendations, to product recommendations. TribeFlow is highly parallel and nearly two orders of magnitude faster than the top state-of-the-art competitors. In order to be application-agnostic we ignore application-specific user and item features, including time-of-day effects, but these can be trivially incorporated into TribeFlow.

To illustrate the performance of TribeFlow consider Figure 1, where we seek to compare the Mean Reciprocal Rank (MRR) of TribeFlow over datasets with up to 1.6 million items and 86 million item visits (further details about this dataset is given in Section 4) against that of state-of-the-art methods such as Multi-core Latent Markov Embedding (MultiLME) [40], personalized ranking LME (PRLME) [13], and Context-aware Ranking with Factorizing Personalized Markov Chains [45] (FPMC). Unfortunately, MultiLME, PRLME, and FPMC cannot finish any of these tasks in less than 10 days while for TribeFlow it takes between one and thirteen hours. In significantly sub-sampled versions of the same datasets we find that TribeFlow is at least 23% more accurate than its competitors.

TribeFlow works by decomposing potentially non-stationary, transient, time-heterogeneous user trajectories into *very* short sequences of random walks on latent environments that are stationary, ergodic, and time-homogeneous. An intuitive way to understand TribeFlow


*WWW 2016,* April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2883059.

# Modeling and Mining Information Popularity Online

Flavio Figueiredo  Jussara M. Almeida
Computer Science Department - Universidade Federal de Minas Gerais, Brazil
{flaviov, jussara}@dcc.ufmg.br

## 1. INTRODUCTION

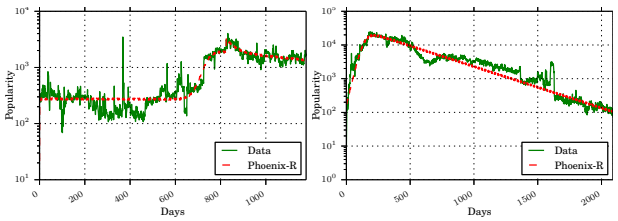
Nowadays, there is an unprecedented amount of user generated content being produced online. This fact is one of the driving forces of what is know as the *social media* phenomenon. Social media shifted how information is produced and propagated. While in traditional media select individuals are responsible for the production, curation and propagation of information, in the social media setting anyone can produce and share information online. One major question in this setting is: *What drives the popularity of information in social media?* This is an interesting question since even with the overload of information that accompanies this mass production of content, some pieces of information manage to attract the attention of millions of users, while the majority remain obscure.

One example of the complexity behind social media popularity is the YouTube channel of *Henri, le Chat-Noir*[1]. The first video of Henri was uploaded in 2007 and remained in obscurity for years. However, in 2012 a user of the Tumblr social network found the video and posted online[2]. Currently, the video and channels has millions of visits from a wide range of different sources (e.g., OSNs, search engines, word-of-mouth and so forth). We can use this single example to motivate our research. Important questions that we raised and approached were as follows:

**What is the impact of incoming links on the popularity of online information?** There are multiple forms through which users can reach content and, thus, there are multiple driving forces that may impact the popularity of information. Identifying such forces is crucial for designing more cost-effective content dissemination strategies. For instance, should a content creator invest time on perfecting the keywords describing content (for better search rankings) or focus on campaigning videos in OSNs? Our current results show that search engines and social propagation inside a service (say YouTube) are major factors in driving popularity [4].

**How does information popularity evolve over time?** Here, we aim at answering if there are different patterns which capture the major trends in which information popularity evolves over time. In a birds-eye-view, we found that there exists a combination of two trends governing the popularity of information. One trend consists of contents that tend to remain attractive over time with an always increasing or steady-state popularity [4]. The other, accounts for content that tend to peak in popularity for a short while, with three different popularity decay characteristics after the peak. Examples of both trends are shown in Figure 1.

**How do users perceive the quality of popular and unpopular information?** Most research in online popularity neglect the users perception of the information being disseminated. We studied if users tend to like very popular content or dislike very unpopular content. Based on a user based study on Mechanical Turk and selected videos from YouTube, we showed that, while users perceptions of content are highly subjective, when the majority of users like a certain YouTube video that video tends to be popular. This was interesting since it shows that there is more than social propagation to popularity. Based on our results [2] we hypothesize that videos like the Henri example would likely be less popular (regardless of OSN propagation) if their content did not appeal to users.




(a) Rock Song  (b) User Dancing Video

**Figure 1: Different YouTube videos as captured by our model.**

**Can we model and predict the future popularity of information?** Two of our most recent results showed that we can model [5] the popularity of information over time and predict the future popularity [1]. Based on our previous findings that we discussed, we developed the Phoenix-R model which can capture the long term popularity evolution as showed in Figure 1. Also, we combined social network propagation and early view patterns of news media to develop prediction models with the user of machine learning tools [1]. These two results show the applicability of our results to mining tasks such as popularity prediction.

With these questions we summarize some of the work on information popularity online that we are pursuing. Our results show that this a promising and new area of research. Currently, we are working on optimizing how early can we predict popularity [3] and practical applications for models and predictions (e.g., search engine rankings or advertising).

[1] `http://www.youtube.com/user/HenriLeChatNoir`
[2] `http://knowyourmeme.com/memes/henri-le-chat-noir`

## 2. REFERENCES

[1] F. Figueiredo, J. Almeida, Y. Matsubara, B. Ribeiro, and C. Faloutsos. Revisit Behavior in Social Media: The Phoenix-R Model and Discoveries. In *Proc. ECML/PKDD*, 2014.

[2] F. Figueiredo, F. Benevenuto, J. Almeida, and K. Gummadi. Does Content Determine Information Popularity in Social Media? A Case Study of YouTube Videos' and their Popularity. In *Proc. CHI.*, 2014.

[3] F. Figueiredo, F. Benevenuto, M. Gonçalves, and J. Almeida. TrendLearner: Early Prediction of Popularity Trends of User Generated Content. *Online at: http://arxiv.org/abs/1402.2351*, 2014.

[4] F. Figueiredo, F. Benevenuto, M. Gonçalves, and J. Almeida. On the Dynamics of Social Media Popularity: A YouTube Case Study. *ACM Transactions on Internet Technology*, To appear.

[5] F. Figueiredo, M. Gonçalves, and J. Almeida. Improving the Effectiveness of Content Popularity Prediction Methods using Time Series Trends. In *ECML/PKDD Discovery Challenge*, 2014.

# Improving the Effectiveness of Content Popularity Prediction Methods using Time Series Trends

Flavio Figueiredo Marcos Gonçalves Jussara M. Almeida
Computer Science Department - Universidade Federal de Minas Gerais, Brazil
{flaviov, mgoncalv, jussara}@dcc.ufmg.br

## ABSTRACT

We here present a simple and effective model to predict the popularity of web content. Our solution, which is the winner of two of the three tasks of the ECML/PKDD 2014 Predictive Analytics Challenge, aims at predicting user engagement metrics, such as number of visits and social network engagement, that a web page will achieve 48 hours after its upload, using only information available in the first hour after upload. Our model is based on two steps. We first use time series clustering techniques to extract common temporal trends of content popularity. Next, we use linear regression models, exploiting as predictors both content features (e.g., numbers of visits and mentions on online social networks) and metrics that capture the distance between the popularity time series to the trends extracted in the first step. We discuss why this model is effective and show its gains over state of the art alternatives.

## Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

## General Terms

Algorithms; Measurement

## Keywords

analytics; predictive; challenge; web page; popularity; host

## 1. INTRODUCTION

With the ever-growing production of online content, characterizing and predicting user engagement (e.g., number of visits or social engagement such as Facebook likes) on web content may have multiple beneficial values such as: (1) understanding the human dynamics of information consumption; (2) supporting the decisions of content producers and providers on different tasks (e.g., marketing and content filtering); and, (3) understanding the physical processes that govern the growth of viewership on the Web. Several previous studies [2–4, 11] have characterized some of the factors that cause the popularity growth of different kinds of web content. Complementarily, various others [1, 6, 7, 11] have focused on the the task of popularity prediction. We focus here on the latter task, aiming at predicting the popularity of a piece of content.

Popularity prediction is a difficult and important task since it mostly translates into income and profits for content providers, creators and consumers alike. For example, more visitors to a web page may lead to more ad-clicks and sales. Moreover, content provisioning to a large amount of users may require decisions such as geographical sharding of content to servers (due to the increased traffic). Thus, if planning is not performed correctly, longer latencies and loading times, and thus, fewer users may be expected. Finally, accurate and early predictions can lead to better services to the end consumer, such as search engine rankings [8].

We here present a simple, and yet effective, model for predicting the popularity of online content. More specifically, we present the winning model of two of the three tasks of the ECML/PKDD 2014 Predictive Analytics Challenge. In the challenge, different features related to the popularity of 30,000 web pages from 100 different hosts were provided. The goal of the challenge is to predict the popularity of 30,000 other pages from the same 100 hosts 48 hours after their upload. The features provided for the task were measured in the first hour after upload for each page.

Our model exploits the temporal features related to web pages (e.g., past visits and social engagement), as well as typical popularity (i.e., number of visits) time series trends which exist in the dataset. Such trends are extracted via unsupervised learning methods. Specifically, it combines the temporal features with features that capture the distances between the popularity time series for each web page and the extracted trends. We present a data characterization that motivates the design of our solution, and show the gains in prediction accuracy (ranging from 15% to 27%) when it is compared to state of the art alternatives.

The rest of this paper is organized as follows. We formally describe the prediction problem and present the state of the art baseline methods in Section 2. In Section 3 we introduce our proposed solution, whereas our experiments and results are presented in Section 4. Finally, Section 5 concludes the paper.

## 2. BACKGROUND

We start this section by defining the content popularity prediction problem (Section 2.1). In this definition, as throughout the rest of the paper, we refer to a particular piece of content as a web page[1]. Next, we discuss existing state of the art solutions used as baselines in our experimental study (Section 2.2).

[1]We here focus on web page popularity prediction, given the goal of the ECML/PKDD Challenge. However, our models are general and can be applied to other types of online content.

ECML/PKDD Discovery Challenge on Predictive Analytics
Nancy, France, September 2014.

# Does Content Determine Information Popularity in Social Media?

## A Case Study of YouTube Videos' Content and their Popularity

**Flavio Figueiredo**[1], **Jussara M. Almeida**[1], **Fabrício Benevenuto**[1]
[1]Department of Computer Science, UFMG, Brazil
{flaviov,jussara,fabricio}@dcc.ufmg.br

**Krishna P. Gummadi**[2]
[2]MPI-SWS, Germany
gummadi@mpi-sws.org

### ABSTRACT

We here investigate *what drives the popularity of information on social media platforms*. Focusing on YouTube, we seek to understand the extent to which content by itself determines a video's popularity. Using mechanical turk as experimental platform, we asked users to evaluate pairs of videos, and compared users' relative perception of the videos' content against their relative popularity reported by YouTube. We found that in most evaluations users could not reach consensus on which video had better content as their perceptions tend to be very subjective. Nevertheless, when consensus was reached, the video with preferred content almost always achieved greater popularity on YouTube, highlighting the importance of content in driving information popularity on social media.

### Author Keywords

Content popularity; social media; user study

### ACM Classification Keywords

H.5.4 Hypertext/Hypermedia: User issues.

### INTRODUCTION

*What drives the popularity of information in social media?* Recently, this question has attracted a lot of research attention as social media sites become increasingly popular. An unresolved part of this question is about the relative roles of two primary forces that drive the popularity of a piece of information: (i) its content, i.e., the interestingness, topicality, or quality of the information *as perceived by users*, and (ii) its dissemination mechanisms, such as propagation by word-of-mouth, blogs or mass media channels. It stands to reason that both factors matter, but the extent to which they impact the popularity of a piece of information remains an open issue.

Many previous studies on how information becomes popular in social media sites focused on dissemination related factors (e.g, social influence, mechanisms that expose content to users, time of upload) [2, 4, 7, 9, 10], ignoring the role of the content itself. Other efforts, instead, analyzed social media content focusing on data mining tasks such as popularity prediction [11] and video classification [5], analyzing popularity differences in content duplicates [2], and exploring content importance as parameter of popularity evolution models [8]. In this paper we take a different and complementary approach, focusing on understanding the extent to which content matters for popularity of videos on YouTube.

Our methodology attempts to assess *users' relative perceptions* of the contents of pairs of videos through user surveys conducted over Amazon mechanical turk. Users in our experiments are exposed only to the video content, and are not subjected to other factors (inherent to the YouTube site) that may impact their perceptions of content (e.g., user comments, social links, appearance of content in external sites). Specifically, we present to users pairs of videos from the same major topic and uploaded around the same date, and ask them to choose which one: (1) *they enjoyed more*, (2) *they would be more willing to share with friends*, and (3) *they predicted would become more popular on YouTube*. These questions target the user's individual perception of content interestingness and of the interests of her social circle (and thus the chance of the content spreading through it), as well as the user's expectations on a global scale. Our goals are to assess, for each of these questions, whether users reach consensus, and, when there is consensus, whether user perceptions match the relative popularity achieved by the videos on YouTube.

We find that users could not reach consensus in many evaluations, even when the popularity (on YouTube) of the evaluated videos differs by orders of magnitude. The lack of consensus is more striking for sharing and liking choices. It also depends on the video topic. This suggests that users' perceptions about content are quite subjective and that content may not be the most important factor that drives popularity in many cases. However, whenever participants reached consensus, their choices mostly match the video with largest popularity on YouTube, suggesting that, in these cases, content has a significant impact and predictive power on video popularity.

The goals of our study complement previous work. In particular, Salganik *et al.* [9] also relied on a user study to understand popularity dynamics. However, they focused on the impact of social influence on popularity, whereas we focus on the role of content and rely on users to evaluate the content in a setup that is isolated (to the extent possible) from dissemination mechanisms that might influence popularity. To our knowledge, the human perceptions of content and how they correlate to popularity in a social media site have not been

**Acknowledgments:** Research supported by InWeb - Institute of Science and Technology for Web Research and by individual grants from CNPq, Capes and Fapemig.


*CHI'14*, April 26–May 1, 2014, Toronto, Canada.

http://dx.doi.org/10.1145/2556288.2557285

# Revisit Behavior in Social Media: The Phoenix-R Model and Discoveries

Flavio Figueiredo[1], Jussara M. Almeida[1],
Yasuko Matsubara[2,3], Bruno Ribeiro[3], and Christos Faloutsos[3]

[1] Department of Computer Science, Universidade Federal de Minas Gerais
[2] Department of Computer Science, Kumamoto University
[3] Department of Computer Science, Carnegie Mellon University

**Abstract.** How many listens will an artist receive on a online radio? How about plays on a YouTube video? How many of these visits are new or returning users? Modeling and mining popularity dynamics of social activity has important implications for researchers, content creators and providers. We here investigate the effect of revisits (successive visits from a single user) on content popularity. Using four datasets of social activity, with up to tens of millions media objects (e.g., YouTube videos, Twitter hashtags or LastFM artists), we show the effect of revisits in the popularity evolution of such objects. Secondly, we propose the PHOENIX-R model which captures the popularity dynamics of individual objects. PHOENIX-R has the desired properties of being: (1) parsimonious, being based on the minimum description length principle, and achieving lower root mean squared error than state-of-the-art baselines; (2) applicable, the model is effective for predicting future popularity values of objects.

## 1 Introduction

*How do we quantify the popularity of a piece of content in social media applications? Should we consider only the audience (unique visitors) or include revisits as well? Can the revisit activity be explored to create more realistic popularity evolution models?* These are important questions in the study of social media popularity. In this paper, we take the first step towards answering them based on four large traces of user activity collected from different social media applications: Twitter, LastFM, and YouTube[4].

Understanding the popularity dynamics of online content is both a challenging task, due to the vast amount and variability of content available, as it can also provide invaluable insights into the behaviors of human consumption [6] and into more effective engineering strategies for online services. A large body of previous work has investigated the popularity dynamics of social media content, focusing mostly on modeling and predicting the *total number of accesses* a piece of content receives [5, 6, 9, 17, 21].

However, a key aspect that has not been explored by most previous work is the effect of revisits on content. The distinction between audience (unique users), revisits (returning users), and popularity (the sum of the previous two) can have large implications for different stakeholders of these applications - from content providers to content

[4] `http://twitter.com` `http://lastfm.com` `http://youtube.com`

# On the Dynamics of Social Media Popularity: A YouTube Case Study

FLAVIO FIGUEIREDO, JUSSARA M. ALMEIDA, MARCOS ANDRÉ GONÇALVES and FABRÍCIO BENEVENUTO, Universidade Federal de Minas Gerais, Brazil

Understanding the factors that impact the popularity dynamics of social media can drive the design of effective information services, besides providing valuable insights to content generators and online advertisers. Taking YouTube as case study, we analyze how video popularity evolves since upload, extracting popularity trends that characterize groups of videos. We also analyze the referrers that lead users to videos, correlating them, features of the video and early popularity measures with the popularity trend and total observed popularity the video will experience. Our findings provide fundamental knowledge about popularity dynamics and its implications for services such as advertising and search.

Categories and Subject Descriptors: H.1.2 [**User/Machine Systems**]: Human Factors

General Terms: Measurement, Human Factors

Additional Key Words and Phrases: youtube, social media, characterization, referrers, popularity growth

**ACM Reference Format:**
Flavio Figueiredo, Jussara M. Almeida, Marcos André Gonçalves, Fabricio Benevenuto, 2014. On the Dynamics of Social Media Popularity: A YouTube Case Study *ACM Trans. Internet Technol.* 1, 1, Article 01 (October 2014), 22 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

## 1. INTRODUCTION

User generated content (UGC) has emerged as the predominant form of online information sharing nowadays. The unprecedented amount of information being produced is one of the driving forces behind the success of the social media phenomenon [Kaplan and Haenlein 2010; Cormode and Krishnamurthy 2008]. This phenomenon is a shift from the traditional media where, instead of content being produced mostly by a few selected individuals, anyone, in theory, can produce and share content online. However, the "information overload" that accompanies the huge amount of social media being produced has its drawbacks. For example, it is ever-so-difficult to find and filter relevant content to oneself. Nevertheless, some pieces of content (or *objects*) succeed in attracting the attention of millions of users, while most remain obscure. This leads to the heavy tailed characteristic of content popularity [Sinha and Pan 2007; Clauset et al. 2009], where a few objects become very popular while most of them attract only a handful of views. *What makes one particular object become hugely popular while the majority receive very little attention? Which factors affect how the popularity of an object will evolve over time?* These are major questions in the social media context that drive our present work.

This research is partially funded by the Brazilian National Institute of Science and Technology for Web Research (MCT/CNPq/INCT Web Grant Number 573871/2008-6), and by the authors' individual grants from Google, CNPq, CAPES and Fapemig. We also thank Caetano Traina, Renato Assunção, Virgilio Almeida, Elizeu Santos-Neto, Matei Ripeanu, and the anonymous reviewers for discussions on drafts of this work.

DOI: http://dx.doi.org/10.1145/0000000.0000000

# On the Prediction of Popularity of Trends and Hits for User Generated Videos

Flavio Figueiredo
flaviov@dcc.ufmg.br
Computer Science Department
Universidade Federal de Minas Gerais

## ABSTRACT

User generated content (UGC) has emerged as the predominant form of media publishing on the Web 2.0. Motivated by the large adoption of video sharing on the Web 2.0, the objective of our work is to understand and predict popularity trends (e.g, will a video be viral?) and hits (e.g, how may views will a video receive?) of user generated videos. Such knowledge is paramount to the effective design of various services including content distribution and advertising. Thus, in this paper we formalize the problem of predicting trends and hits in user generated videos. Also, we describe our research methodology on approaching this problem. To the best of knowledge, our work is novel in focusing on the problem of predicting popularity trends complementary to hits. Moreover, we intend on evaluating efficacy of our results not only based on common statistical error metrics, but also on the possible online advertising revenues our predictions can generate. After describing our proposal, we here summarize our latest findings regarding (1) uncovering common popularity trends; (2) measuring associations between UGC features and popularity trends; and (3) assessing the effectiveness of models for predicting popularity trends.

## Categories and Subject Descriptors

C.4 [**Computer Systems Organization**]: Performance of Systems—*Measurement techniques*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

## Keywords

UGC; video; popularity; trends

## General Terms

Human Factors; Design; Measurement

## 1. INTRODUCTION

On the Web 2.0, user generated content (UGC) has become the de-facto form of media publishing on some of the most popular Internet applications nowadays [6]. Focusing on video content, websites such as YouTube [1] receive over 800 million unique users monthly, attracting over 1 million different advertisers [23]. Even niche applications, such as Vimeo [2], which focuses on independent filmmakers, manage to attract over 70 million unique users monthly [19].

Given the success of such applications and the current large volume of videos consumed daily, understanding how users find such content and how content popularity evolves provide valuable insights for content generators, online advertisers and Internet service providers (ISPs), amongst others. For instance, from a systems perspective, understanding these properties may drive the design of better analytic tools, a major market segment nowadays. Online advertisers may also benefit from this information to better place contextual advertisements, while ISPs could exploit it to develop more cost-effective content delivery platforms and caching systems. From a social perspective, understanding the properties of video popularity trends could be used to better comprehend the human dynamics of consumption processes [7]. Also, content producers could use insights on how user collaboration and collaborative social activities on Web 2.0 applications may impact content popularity, providing information on aspects related to their own fame on video sharing applications.

Most previous efforts, which are focused on predicting the popularity of a piece of content measured at a specific future date [16–18, 22], are still preliminary, as they provide limited knowledge on which features and system mechanisms (e.g., search, related videos, etc) contribute the most to popularity growth. Analyzing the importance of such features to popularity growth is key to provide scalable alternatives to service design, as solutions based on content analysis are less scalable in (user generated) videos. Moreover, there is little effort towards predicting popularity evolution (or trends), which may also provide valuable knowledge. For instance, online advertisers and content delivery systems could benefit more from predicting not only a final popularity measure for UGC, but also whether its popularity trend is increasing and how stable it is likely to be over time.

*In sum, our proposed research aims at understanding the importance and utility of various features, particularly referrers (i.e, incoming links to videos), on the popularity evolution of individual user generated videos and exploiting them to develop methods to predict future popularity measures and trends of those videos.*

The rest of this paper is organized as follows. Section 2 describes our problem statement and research goals. We describe our current methodology on addressing our goals on Section 3. The current state of our research is described in Section 4 while our related work is addressed in Section 5. Section 6 concludes this paper.

[1] `http://youtube.com`
[2] `http://vimeo.com`


*WSDM'13,* February 4–8, 2013, Rome, Italy.

# The Tube over Time: Characterizing Popularity Growth of YouTube Videos

Flavio Figueiredo Fabrício Benevenuto Jussara M. Almeida
{flaviov, fabricio, jussara}@dcc.ufmg.br
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte - Brazil

## ABSTRACT

Understanding content popularity growth is of great importance to Internet service providers, content creators and online marketers. In this work, we characterize the growth patterns of video popularity on the currently most popular video sharing application, namely YouTube. Using newly provided data by the application, we analyze how the popularity of individual videos evolves since the video's upload time. Moreover, addressing a key aspect that has been mostly overlooked by previous work, we characterize the types of the referrers that most often attracted users to each video, aiming at shedding some light into the mechanisms (e.g., searching or external linking) that often drive users towards a video, and thus contribute to popularity growth. Our analyses are performed separately for three video datasets, namely, videos that appear in the YouTube top lists, videos removed from the system due to copyright violation, and videos selected according to random queries submitted to YouTube's search engine. Our results show that popularity growth patterns depend on the video dataset. In particular, copyright protected videos tend to get most of their views much earlier in their lifetimes, often exhibiting a popularity growth characterized by a viral epidemic-like propagation process. In contrast, videos in the top lists tend to experience sudden significant bursts of popularity. We also show that not only search but also other YouTube internal mechanisms play important roles to attract users to videos in all three datasets.

## Categories and Subject Descriptors

C.4 [**Computer Systems Organization**]: Performance of Systems—*Measurement techniques*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

## General Terms

Human Factors, Measurement


*WSDM'11,* February 9–12, 2011, Hong Kong, China.


## Keywords

YouTube, video popularity, popularity growth, referrer

## 1. INTRODUCTION

Understanding content popularity growth on the Internet is of great relevance to a broad range of services, from technological, economical and social perspectives. Such understanding can drive the design of cost-effective caching and content distribution mechanisms as well as uncover potential bottlenecks in system components such as search engines [6]. Moreover, predicting popularity is also important not only for supporting online and viral marketing strategies as well as effective information services (e.g., content recommendation and searching services) [12] but also because it may uncover new (online and offline) business opportunities. From a sociological point of view, a deep study of popularity evolution may also reveal properties and rules governing collective user behavior [10].

Online Social Networks (OSNs) are currently a major segment of the Internet. Considering video sharing OSNs, YouTube[1] is the one with the largest number of registered users [1], who upload and share their videos at a staggering rate. Indeed, it has been reported that the amount of content uploaded to YouTube in 60 days is equivalent to the content that would have been broadcasted for 60 years, without interruption, by NBC, CBS and ABC altogether [2]. Moreover, YouTube has reportedly served over 100 million users only on January 2009 [1], with a video upload rate equivalent to 10 hours per minute [2]. At such unprecedented user and content growth rates, understanding video popularity on YouTube becomes a challenge of utmost importance, as the myriad of different contents make user behavior and attention span highly variable and unpredictable [6].

As argued by Willinger *et al.* [20], most previous analyses of OSNs have treated such systems as static. Most of them focus on analyzing structural properties of single snapshots of relationship networks (e.g., friendship network) that emerge in such systems [3, 5, 15]. However, since OSNs are inherently dynamic, these studies fail to address key properties of the underlying system dynamics. Regarding one such property, namely popularity, a few studies have analyzed YouTube with respect to video popularity characteristics [6, 9, 10] and prediction [14, 19]. However, most of them, despite covering a rich set of popularity properties and their implications for system design, focused on only a single or

[1] `http://www.youtube.com`
[2] `http://www.youtube.com/t/fact_sheet`

# Content Popularity Evolution in Online Social Networks

Flavio Figueiredo Fabricio Benevenuto Jussara M. Almeida
{flaviov, fabricio, jussara}@dcc.ufmg.br
Universidade Federal de Minas Gerais (UFMG)

## ABSTRACT

Understanding content popularity growth on Online Social Networks (OSNs) is of great importance to Internet service providers, content creators and online marketers. However, most previous studies of OSNs are based on static views of the system, thus neglecting the temporal evolution of the network, and a possible correlation with content popularity growth. Moreover, previous analyses also greatly neglect the impact of the referrers (i.e., incoming links from external sites) on content popularity. We here provide some initial results on the analysis of content popularity growth in YouTube videos. Our study is based on three video datasets, namely popular videos, randomly collected videos, and copyright protected videos, with distinct characteristics in terms of temporal popularity evolution. We also characterize the different referrers that most often lead users to YouTube videos. Our results shed some light into aspects that impact content popularity growth.

## Categories and Subject Descriptors

C.4 [**Computer Systems Organization**]: Performance of Systems—*Measurement techniques*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

## General Terms

Human Factors, Measurement

## Keywords

OSNs, YouTube, video popularity, popularity growth

## 1. THEORETICAL BACKGROUND

This paper describes a PhD work, being developed at the Universidade Federal de Minas Gerais. The work started in July 2010 and is expected to finish by July 2015.

Given that Online Social Networks (OSNs) are currently a major segment of the Internet, understanding content popularity growth on these networks is of great relevance to a broad range of services, from technological, economical and social perspectives. Such understanding can drive the design of cost-effective caching and content distribution mechanisms as well as uncover potential bottlenecks in system components such as search engines [5]. Moreover, predicting popularity is also important not only for supporting online and viral marketing strategies as well as effective information services (e.g., content recommendation and searching services) but also because it may uncover new (online and offline) business opportunities. From a sociological point of view, a deep study of popularity evolution may also reveal properties and rules governing collective user behavior [6].


*Webmedia-WTD'11,* October 3–6, 2011, Florianopolis, Brasil.


## 2. OBJECTIVES

The main objective of our work is to understand the diffusion and evolution of content popularity in large scale OSNs. In particular, we are interested on dealing with OSNs which focus on user created content (UGC) [1], due to the volume [5, 6] and more complicated nature of such media [4]. One representative example of such OSNs is YouTube[2], being the largest video sharing network nowadays.

In broader terms, we aim at understanding the evolution of content popularity with respect to three main research challenges (RC): (1) popularity growth patterns, which are related to the different patterns of popularity evolution across UGC content; (2) the referrers (i.e. incoming links) of UGC, which deals with how users find content on OSN and how this impacts popularity evolution; and finally, (3) how changes in the structure of the OSN affect popularity.

We begin our study with a review of the related literature in Section 3. A description of each challenge is presented in Section 4, while our research methodologies are presented in Section 5. In order to provide initial insights on *RC 1-2*, we characterized the growth patterns of video popularity on YouTube [8]. Using newly provided data by the application, we analyzed how the popularity of individual videos evolves since the video's upload time. We also characterize the different referrers for each video. Our results reveal differences in popularity evolution patterns depending on different video samples (top, random and copyrighted). These are presented in Section 6. Section 7 concludes the work.

## 3. RELATED WORK

**Static views of popularity:** There have been a few studies that address content popularity on OSNs, and, particularly, on video sharing systems. Cha *et al.* [5] presented

[1]Online radios, such as LastFM (`http://www.last.fm`), are examples OSNs which does not deal with UGC.
[2]`http://www.youtube.com`